2012

Object Detection and Recognition for Visually Impaired People

Shuihua Wang
CUNY City College

Object Detection and Recognition for Visually Impaired People

Thesis

Submitted in partial fulfillment of

the requirement for the degree

Master of Engineering (Electrical Engineering)

at

The City College of New York

of

The City University of New York

by

Shuihua Wang

May 2012

Approved:

Professor Yingli Tian, Thesis advisor

Professor Roger Dorsinville, Chairman

Department of Electrical Engineering

Table of Contents

Figure Lists

Table Lists

Acknowledgment

I would like to express my gratitude to my supervisor, Professor Yingli Tian, whose expertise, understanding, and patience, added considerably to my graduate experience. I appreciate her vast knowledge and skill in many areas, and her assistance in writing papers and this thesis.

I would like to thank the other members in the CCNY Media Lab, Shizhi Chen, Xiaodong Yang, Chuchai Yi, Chenyang Zhang, for the assistance they provided at all levels of the research project.

This work was supported by NIH 1R21EY020990, NSF IIS-0957016 and EFRI-1137172, DTFH61-12-H-00002, ARO W911NF-09-1-0565, Microsoft Research, and CITY SEEDs grant.

I would also like to thank my family and my parents for the support they provided me through my entire life and in particular, I must acknowledge my husband, he encouraged me go through my graduate life and come to this lab. Finally, I would like to express my great appreciation to my 7 months old infants, his coming makes me more encouraged.

Abstract

Object detection plays a very important role in many applications such as image retrieval, surveillance, robot navigation, wayfinding, etc. In this thesis, we propose different approaches to detect indoor signage, stairs and pedestrians.

In the first chapter we introduce some related work in this field.

In the second chapter, we introduced a new method to detect the indoor signage to help blind people find their destination in unfamiliar environments. Our method first extracts the attended areas by using a saliency map. Then the signage is detected in the attended areas by using bipartite graph matching. The proposed method can handle multiple signage detection. Experimental results on our collected indoor signage dataset demonstrate the effectiveness and efficiency of our proposed method. Furthermore, saliency maps could eliminate the interference information and improve the accuracy of the detection results.

In the third chapter, we present a novel camera-based approach to automatically detect and recognize restroom signage from surrounding environments. Our method first extracts the attended areas which may content signage based on shape detection. Then, Scale-Invariant Feature Transform (SIFT) is applied to extract local features in the detected attended areas. Finally, signage is detected and recognized as the regions with the SIFT matching scores larger than a threshold. The proposed method can handle multiple signage detection. Experimental results on our collected restroom signage dataset demonstrate the effectiveness and efficiency of our proposed method.

In the fourth chapter, we develop a new framework to detect and recognize stairs and pedestrian crosswalks using a RGBD camera. Since both stairs and pedestrian crosswalks are featured by a group of parallel lines, we first apply Hough transform to extract the concurrent parallel lines based on the RGB channels. Then, the Depth channel is employed to further recognize pedestrian crosswalks, upstairs, and downstairs using support vector machine (SVM) classifiers. Furthermore, we estimate the distance between the camera and stairs for the blind users. The detection and recognition results on our collected dataset demonstrate that the effectiveness and efficiency of our proposed framework

Keywords: Blind people, Navigation and wayfinding, Camera, Signage detection and recognition, Independent travel.

Chapter 1: Background

Based on the study of the World Health Organization (WHO), there were about 161 million visually impaired people around the world in 2002, about 2.6% of the total population. Among these statistics, 124 million had low vision and 37 million were blind [1]. Robust and efficient indoor object detection can help people with severe vision impairment to independently access unfamiliar indoor environments and avoid dangers [2] . Context information (including signage and other visual information) plays an important role in navigation and wayfinding for unsighted persons. As shown in Figure 1, signage is particularly important for discriminating between similar objects in indoor environments such as elevators, bathrooms, exits, and office doors.

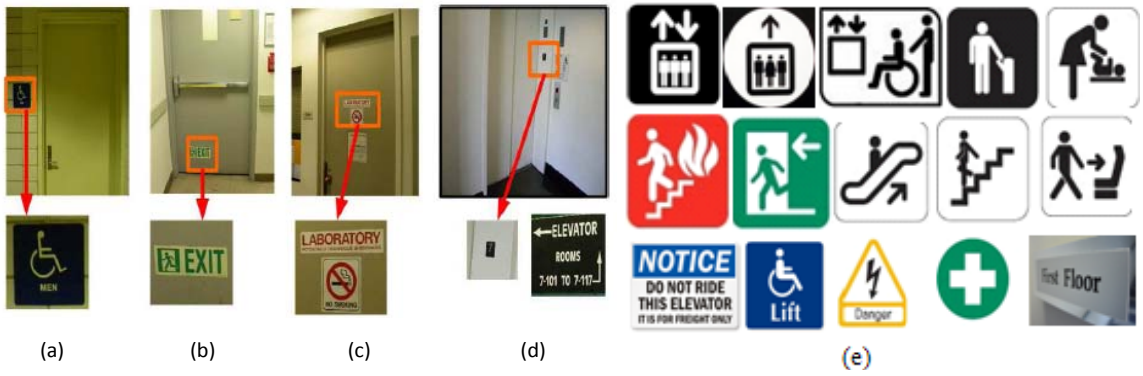


Figure 1. Typical indoor signage: (a) a bathroom, (b) an exit, (c) a laboratory, (d) an elevator, (e) stairs

Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of a priori objects of a certain class (such as faces, signs, buildings, etc) in digital images and videos captured by cameras [3]. Camera-based indoor signage detection is a challenging problem due to the following factors: 1) large variations of appearance and design (shape, color, texture, etc.) of signage in different buildings; and 2) large variations in the camera view and image resolution of signage due to changes in position and distance between the blind user with wearable cameras and the targeted signage. Some examples of camera used in our experiment are shown in Figure 2.

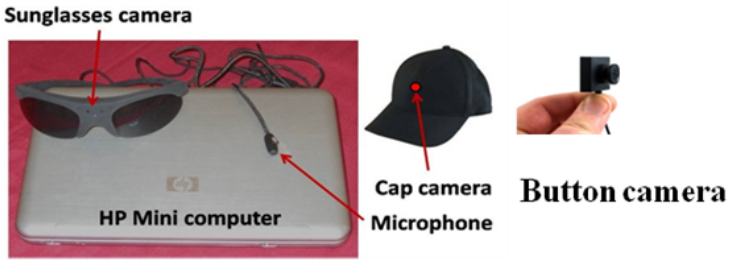


Figure 2. Wearable camera used in our test procedures

Object detection and recognition is a fundamental component for scene understanding. The human visual system is powerful, selective, robust, and fast [4]. It is not only very selective, which allows us to distinguish among very similar objects, such as the faces of identical twins, but also robust enough to classify same category objects with large variances (e.g. changes of position, scale, rotation, illumination, color, occlusion and many other properties). Research shows that the human visual system can discriminate among at least tens of thousands of different object categories [5]. Object recognition processes in the human visual system are also very fast: it can take as little as 100 to 200 ms [6][7][8]. However, it is extremely difficult to build robust and selective computer vision algorithms for object recognition which can handle very similar objects or objects with large variations. For example, state-of-the-art object detection methods require hundreds or thousands of training examples and very long durations to learn visual models of one object category [9][10][11].

Many disability and assistive technologies have been developed to assist people who are blind or visually impaired. The vOICe vision technology for the totally blind offers sophisticated image-to-sound renderings by using a live camera [12]. The Smith-Kettlewell Eye Research Institute developed a series of camera phone-based technological tools and methods for the understanding, assessment, and rehabilitation of blindness and visual impairment [13][14][15][16][17], such as text detection [17], crosswatch [14] [16], and wayfinding [15]. To help the visually impaired, Zandifar *et al.* [18] used one head-mounted camera together with existing OCR techniques to detect and recognize text in the environment and then convert the text to speech. Everingham *et al.* [19] developed a wearable mobility aid for people with low vision using scene classification in a Markov random field model framework. They segmented an outdoor scene based on color information and then classified the regions of sky, road, buildings etc. Shoval *et al.* [20] discussed the use of mobile robotics technology in the GuideCane device, a wheeled device pushed ahead of the user via an attached cane for the blind to avoid obstacles. When the GuideCane detects an obstacle it steers around it. The user immediately feels this steering action and can follow the GuideCane's new path. Pradeep *et al.* [21]describes a stereo-vision based algorithm that estimates the underlying planar geometry of the 3D scene to generate hypotheses for the presence of steps. The Media Lab at the City College of New York has been developed a number of computer vision based technologies to help blind people including banknote recognition [22], clothing pattern matching and recognition [23][24], text extract [25][26][27], and navigation and wayfinding [28][29][30]. Coughlan *et al.* [31] developed a method of finding crosswalks based on figure-ground segmentation, which they casted in a graphical model framework for grouping geometric features into a coherent structure. Advanyi *et al*. [32] employed the Bionic eyeglasses to provide the blind or visually impaired individuals the navigation and orientation information based on an enhanced color preprocessing through mean shift segmentation. Then detection of pedestrian crosswalks was carried out via a partially adaptive Cellular Nanoscale Networks（CNN）algorithm. Se *et al*. [34] proposed a Gabor filter based texture detection method to detect distant stair cases. When the stairs are close enough, stair cases were then detected by looking for groups of concurrent lines, where convex and concave edges were portioned using intensity variation information. Stair cases pose was also estimated by homograph search model. Se *et al* [33] further extended the method to detect zebra crosswalks. They first detected the crossing lines by looking for groups of concurrent lines. Edges were then partitioned using intensity variation information. Uddin *et al.* [35] proposed a bipolarity-based segmentation and projective invariant-based method to detect zebra crosswalks. They first segmented the image on the basis of bipolarity and selected the candidates on the basis of area, then extracted feature points on the candidate area based on the fisher criterion. The authors recognized zebra crosswalks based on the projective invariants. Lausser *et al* [36] introduced a visual zebra crossing detector based on the Viola-Jones approach. Tian *et al*.[36]developed a proof-of-concept computer vision-based wayfinding aid for blind people to independently access unfamiliar indoor environments. They mainly focus on indoor object detection and context information extraction and recognition.

Although many efforts have been made, how to apply this vision technology to help blind people understand their surroundings is still an open question.

The rest of the thesis is organized as following way: Chapter 2 demonstrates indoor signage detection based on saliency map and Bipartite Graph Matching, chapter 3 is Camera-based Signage Detection and Recognition for Blind Persons chapter 4 is about stair cases and pedestrian Detecting Stairs and Pedestrian Crosswalks for the Blind by RGBD Camera.

Chapter 2: Indoor Signage Detection Based on Saliency Map and Bipartite Graph Matching

2.1. Introduction

Scholars tend to combine the window-sliding technique with the classifier to detect regions of an image at all locations and scales that contain the given objects. However, the window sliding method suffers from two shortcomings [38]: 1) high processing time; and 2) inaccuracy of detection results due to different background. Therefore, we propose a new method to detect indoor signage which first employs the saliency map [39][40] to extract attended areas and then applies bipartite graphic matching [41][42][43] to recognize indoor signage only at the attended areas instead of the whole image, which can increase the accuracy and reduce the computation cost.

Detection of indoor signage can help blind people to find their destinations in unfamiliar environments. In this paper, we propose a new method to detect indoor signage by combining saliency map based attended area extraction and bipartite graph matching based signage recognition. The paper is organized as following: Section 2.2 describes the methodology of our proposed algorithm, including 1) calculating the saliency map to detect attended area, and 2) the bipartite graph matching for pattern recognition. Section 2.4 displays our experimental results and demonstrates the effectiveness and efficiency of the proposed algorithm. Section 2.5 concludes the paper.

2.2. Method of Indoor Signage Detection

Our method consists of two phases as shown in Figure 3. In the first phase, attended areas are detected via saliency map based on color, intensity, and orientation. Then, the scaled patterns are detected within attended areas using bipartite graph matching. To localize the patterns, a window sliding method is employed to search the attended areas.

2.2.1 Build Saliency Maps

Saliency maps are used to represent the conspicuity at every location in the visual field by a scalar quantity and to guide the selection of attended locations based on the spatial distribution of saliency [44]. In analogy to the center-surround representations of elementary visual features, bottom-up saliency is thus determined by how different a stimulus is from its surround, in many sub-modalities and at many scales [45]. Saliency at a given location is determined primarily by how different this location is from its surroundings in color, orientation, motion, depth, etc.

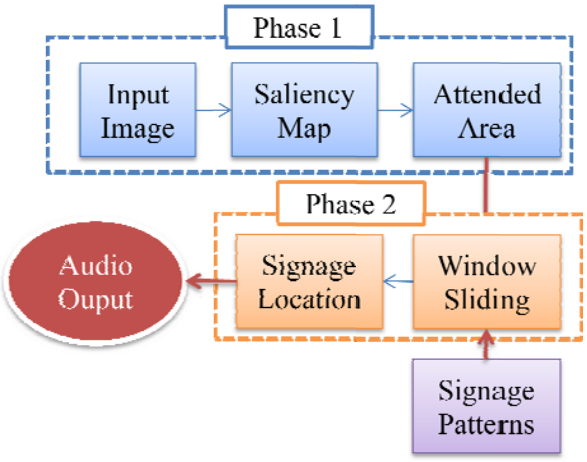


Figure 3. Flow chart of the proposed algorithm

Figure 4. Architecture of building saliency maps

As shown in Figure 4, the different visual features that contribute to attentive selection of a stimulus (color, intensity, and orientation) are combined into one saliency map. The saliency map which integrates the normalized information from the individual feature maps into one global measure of conspicuity. The detailed procedures are discussed in following sections.

2.2.2 Initialization

We first resize the input color images at 640 × 480 resolution. Then, for each image, nine spatial scales are created using dyadic Gaussian pyramids, which progressively low-pass filter and sub-sample the input image, yielding horizontal and vertical image-reduction factors ranging from 1:1 to 1:256 in eight octaves.

We compute saliency maps based on intensity, color, and orientation for each pixel by linear 'center-surrounded' operations similar to visual receptive fields. Center-surround feature extraction is implemented in the function of the difference between fine and coarser scales; the center is a pixel at the scales $c \in \{2,3,4\}$, and the surround is a pixel at scale $s=c+\delta$, with $\delta \in \{3,4\}$. The across-scale difference between two maps is gained by interpolation to the finer scale and point-by-point subtraction. Using several scales for both c and $\delta = s-c$ yield truly multi-scale feature extraction, by including different size ratios between the centers and surround regions.

2.2.3 Intensity-based Saliency Map

With r, g, and b being the red, green, and blue channels of the input image, respectively, an intensity image I is achieved as $I = (r+g+b)/3$. Here I is used to create a Gaussian pyramid $I(\sigma)$, where $\sigma \in \{0,1,2,...8\}$ denotes the scale. The r, g, and b channels are normalized by I in order to decouple hue from intensity. However, the hue variation are not perceivable at very low luminance (and hence are not salient), so normalization is only applied at the locations where I is larger than 10% of its maximum over the entire image and other locations yield zero. Four broadly-tuned color channels are created: $R = r-(g+b)/2$ for red, $G = g-(r+b)/2$ for green. $B = b-(r+g)/2$ for blue, and $Y = (r+g)/2-|r-g|/2-b$ for yellow, and negative values are set to zero for the R, G, B and Y values gained from above equations, 4 Gaussian pyramids $R(\sigma)$, $G(\sigma)$, $B(\sigma)$, and $Y(\sigma)$ are created from these color channels.

Center-surround differences between a 'center' fine scale c and a 'surround' coarser scale s yield the feature maps. The first group of feature maps is concerned with intensity contrast, which is detected by neurons

sensitive either to dark centers on bright surrounds or to bright centers on dark surrounds. In this paper, both types of sensitivities are simultaneously computed by using a rectification in six maps $I(c,s)$.

$$I(c,s) = |\, I(c) - I(s) \,| \tag{1}$$

2.2.4 Color-based Saliency Map

The second group of saliency maps is similarly built for the color channels, which in cortex are represented by a so-called "color double-opponent" system: in the center of their receptive fields, neurons are excited by one color such as red and inhibited by another such as green, while the converse is true in the surround [46]. Such spatial and chromatic opponency exists for the red/green, green/red, blue/yellow, and yellow/blue color pairs in human primary visual cortex. Accordingly, maps $RG(c,s)$ are created in the function to simultaneously account for red/green and green/red double opponency (2) and $BY(c,s)$ for blue/yellow and yellow/blue double opponency (3)

$$RG(c,s) = |\, (R(c) - G(c)) - (G(s) - R(s)) \,| \tag{2}$$
$$BY(c,s) = |\, (B(c) - Y(c)) - (Y(s) - B(s) \,| \tag{3}$$

2.2.5 Orientation-based Saliency Map

Local orientation information is gained from I by oriented Gabor pyramids $O(\sigma, \theta)$ where $\sigma \in [0, 1, \ldots, 8]$ represents the scale and $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ stands for the preferred orientation. Gabor filters are the product of a cosine grating and a 2D Gaussian envelope, approximating the receptive field sensitivity profile of orientation-selective neurons in primary visual cortex. Orientation feature maps $O(s, c, \theta)$ encode local orientation contrast between the centers and surround scales.

$$O(c,s,\theta) = |\, O(c,\theta) - O(s,\theta) \,| \tag{4}$$

In total, 42 feature maps are created: 6 from intensity, 12 from color, and 24 from orientation.

2.2.6 Combination of Saliency Maps

The difficulty in combining different maps is that they represent a priori not comparable modalities, with different dynamic ranges and extraction mechanisms [47]. Furthermore, because all 42 maps are combined, salient objects which are strong in only a few maps may be masked by noise or by less-salient objects present in a larger number of maps.

Because of the absence of top-down supervision, a map normalization operator $N(.)$ is proposed, which globally promotes maps in which a small number of strong peaks of activity is present, meanwhile globally suppressing maps which contain numerous comparable peak responses. $N(.)$ consists of:

(1) Normalizing the values range in the map to a fixed range [0, …, M], in order to eliminate modality dependent amplitude differences;
(2) Finding the location of the map's global maximum M and computing the average $\overline{m}$ of all its other local maxima;
(3) Globally multiplying the map by $(M - \overline{m})^2$.

Only local maxima activities are considered, such that $N(.)$ compares responses associated with meaningful 'active spots' in the map and ignores homogeneous areas. Comparing the maximum activity in the whole map to the average overall activity measures the difference between the most active location and the

average. If the difference is large, the most active location stands out, and the map is strongly promoted. Otherwise, the map contains nothing unique and is suppressed.

Feature maps are combined into three "conspicuity maps", $\bar{I}$ for intensity (5), $\bar{C}$ for color (6), and $\bar{O}$ for orientation (7), at the scale $\sigma = 4$ of the saliency map. They are gained by across-scale addition $\oplus$ which consists of reduction of each map to scale four and point-by-point addition.

$$\bar{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} N(I(c,s)) \qquad (5)$$

$$\bar{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} N(\mathrm{RG}(c,s)) + N(BY(c,s))] \qquad (6)$$

For orientation, 4 intermediary maps are created by combining the six maps for a given θ and then are combined into a single orientation conspicuity map:

$$\bar{O} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} N(I(c,s)) \qquad (7)$$

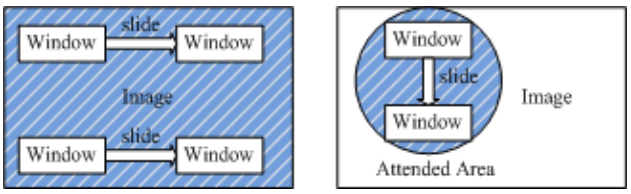
The motivation of the creation of three separate channels ($\bar{I}$, $\bar{C}$, and $\bar{O}$) is based on the hypothesis that similar features compete strongly for saliency, meanwhile different modalities contribute separately to the saliency map. The three conspicuity maps are normalized and grouped into the final input *S* to the saliency map.

$$S = \frac{1}{3}\left(N(\bar{I}) + N(\bar{C}) + N(\bar{O})\right) \qquad (8)$$

From above equations, we get the most salient image locations based on the maximum of the saliency map. We can simply detect the attended areas as the connected points by comparing the value of the saliency map and a threshold.

2.3. Bipartite Graph Matching Based Indoor Signage Detection

2.3.1. Detecting Signage in Attended Areas



(a) (b)

Figure 5. Applying window sliding method in attended areas improves computation efficiency: (a) Traditional (b) Our approach.

Figure 5 shows the concept of the window sliding method [48], which is a popular technique for identifying and localizing objects in an image. The traditional approach involves scanning the whole image as the shaded area in Figure 5(a) with a fixed-size rectangular window and applying a classifier to the sub-image defined by the window. The classifier extracts image features within the window and returns the probability that the window bounds a particular object. The process is repeated on different scales so that objects can be detected at any size [49]. Usually non-maximal neighborhood suppression is applied to the output to remove multiple detections of the same object. In order to improve the efficiency, our method scans the attended areas as the shaded circle in Figure 5 (b). Our method reduces the scan areas, which helps to reduce the sliding window algorithm's processing time.

2.3.2. Bipartite Graph Matching

To detect indoor signage, we employ the "Bipartite Graph Matching (BGM)" algorithm to compare the query patterns and the slide-windows at the attended areas. The detailed procedures of the BGM algorithm can be found at papers [50] [51]. Suppose A denotes the window and B denotes the sub-image covered by window A. The BGM algorithm first calculate the edges of the window and the sub-image, followed by evaluating the degree of the match as

$$BGM(A,B) = \frac{\omega}{N} \tag{9}$$

Where ω denotes the sum of the pixels that exist in the edge images of both A and B, and N denotes the size of the window. Larger values of function (9) correspond to a better match; therefore, the BGM algorithm scans the image and finds the location with the largest matching score:

$$B^* = \arg\max\left(BGM(A,B)\right) \tag{10}$$

The maximization can be solved by a gradient-based optimization technique. However, the function (10) is non-convex and multi-modal, so the gradient-based optimization technique is easy to mislead in local extreme. In this paper, an exhaustive searching method is employed to find the global maxima.

2.3.3. Generalization to Multi-pattern Detection

Our method can handle both single signage and multiple signage detection. The pseudo-codes of multi-pattern detection are listed below, which is based on multi runs of aforementioned single pattern detection described above. Here, I denotes original image; P_i denotes ith pattern; S denotes the saliency map; BGM denotes the value of bipartite graph matching; L_i denotes the location found by ith pattern.

Step 1 *Initialization. Input I, P_1, P_2, ... P_N;*
Step 2 *S=GetSaliencyMap(I);*
Step 3 *for i=1: N*
[BGM_i, L_i]=SinglePatternDetection(I, S, P_i);
end
Step 4 *i*=argmax {BGM_i} ;*
Step 5 *Output i* and L_{i^*}.*

2.4. Experiments and Discussion

The experiments are carried on the Windows XP operation system with 2GB Hz processor and 1GB memory. We also developed an in-house GUI as a friendly interface between human and computers. The GUI can run on any computer with Matlab.

2.4.1. Single Pattern Detection

Independent travel is well known to present significant challenges for individuals with severe vision impairment, thereby reducing quality of life and compromising safety. Based on our survey with blind users, detection of indoor signage such as elevator buttons and restrooms has high priority. Therefore, our experiments focus on the detection of signage.

Figure 6 displays detection of elevator buttons. The camera-captured image of six different elevator buttons is shown in Figure 6(a). Figure 6(b) is the corresponding saliency map extracted from the original image by using intensity, color, and orientation. The bright pixels indicate the attended areas. Figure 6(c) and Figure 6(e) show the query symbols of "open" and "close" buttons which are recorded in the query database of indoor signage. Figure 6(d) and Figure 6(f) demonstrate the final detection results (the red rectangular boxes).

Similarly, the detection of restroom signage is displayed in Figure 7 as (a) shows the original image including the signage of both “Women’s” and “Men’s” restrooms; (b) shows the saliency map of the original image where the bright regions indicate the attended areas; (c) and (e) are the query patterns; (d) and (f) are the final detection signage results of the “Men’s” and “Women’s” restroom.

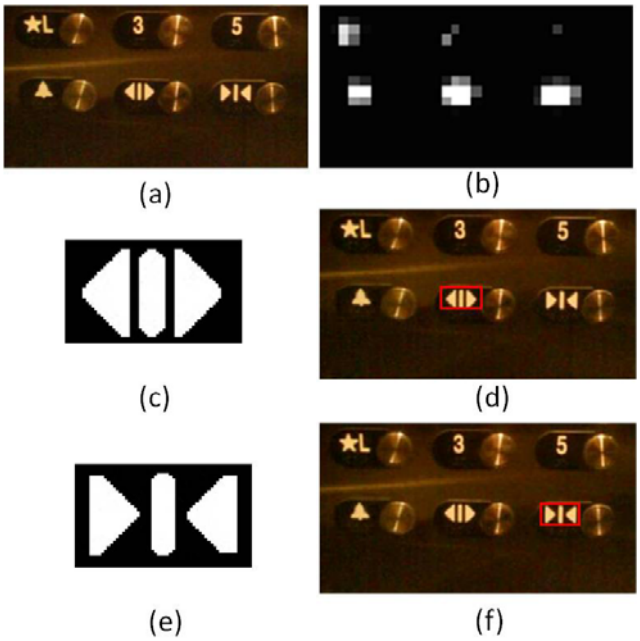


Figure 6. Elevator button detection: (a) Original image; (b) Saliency map; (c) Query pattern of “Open” symbol; (d) Detected “Open” symbol; (e) Query pattern of “Close” symbol; and (f) Detected “Close” symbol.

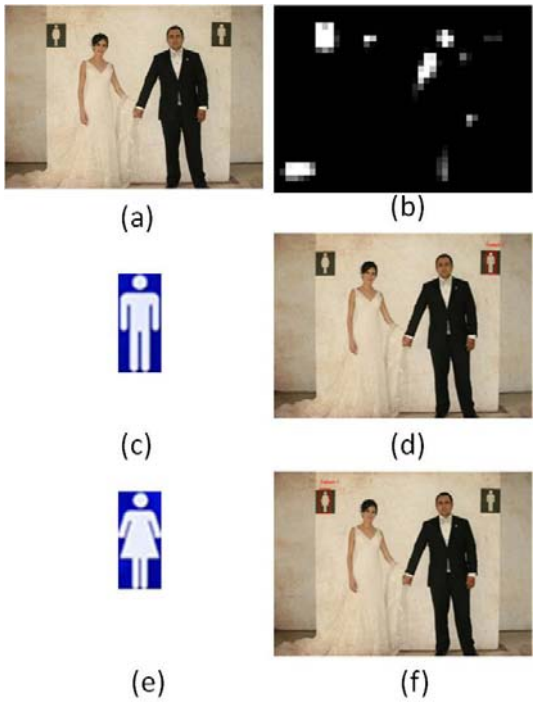


Figure 7. Restroom signage detection: (a) Original image; (b) Saliency map; (c) Query pattern of “Men’s” signage; (d) Detected “Men’s” signage; (e) Query pattern of “Women’s” signage; and (f) Detected ‘Women’s” signage.

Figure 8. Multi-pattern detection results. First row: Multiple query patterns; Second row: Original image, Saliency map, and the detected result of pattern 2 (signage of Men's restroom); Third row: Original image, Saliency map, and the detected result of pattern 1 (signage of Women's restroom). Fourth row: riginal image, Saliency map, and the detected result of pattern 3 (signage of Disabled restroom).

2.4.2. Multi-Pattern Detection

We further extend our algorithm to detect multiple patterns. In this case, the blind user will give multiple query patterns. As shown in Figure 9 (first row), the query patterns include both "Men's" and "Women's" restroom. We need to detect the patterns as well as recognize which pattern is found. The original image, corresponding saliency map, and the detected signage are shown in (the second row) for a "Men's" and a "Women's" restroom (the third row) respectively.

We further evaluate the multi-pattern method to detect the four emergency exit signage of up, down, left, and right directions. The detection results are demonstrated in Figure 9.

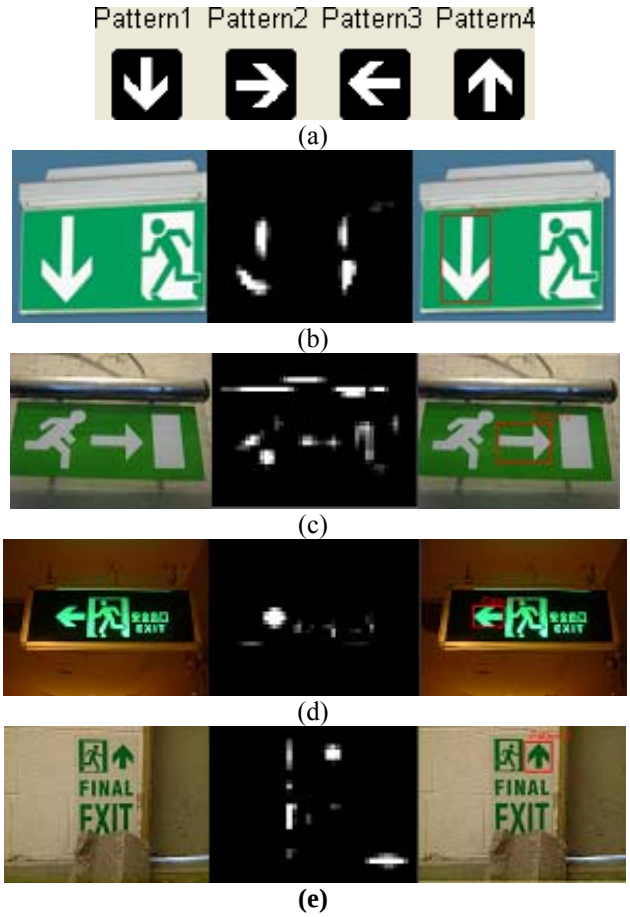


Figure 9. Emergency exit signage detection: (a) Query patterns of direction symbols; (b) original image, saliency map, and detection result of the "Down" symbol; (c) original image, saliency map, and detection result of the "Right" symbol; (d) original image, saliency map, and detection result of the "Left" symbol; and (e) original image, saliency map, and detection result of the "Up" symbol.

2.4.3. Experimental Results

Our proposed method is evaluated by a database of 162 indoor signage images with 8 different types of signage including restrooms (Men's and Women's), elevator buttons (Open and Close), and exit directions (left, right, up, and down). Some examples are shown in Figure 9 which contains variations in lighting, resolution, and camera angle.

In single-pattern detection, we use 40 restroom sign images (20 women's and 20 men's) and 40 elevator button images (20 close and 20 open). We correctly detected 18 "Women's" signs, 16 'Men's' signs, 17 "Close" signs, and 17" open" signs as shown in Table 1

In multi-pattern detection, we use 82 direction signage (20 up, 20 down, 20 left, and 22 right) and 40 restroom signage (20 women and 20 men). In this experiment, 14 "Women's" signs, 16 'Men's' signs, 18 "Up" directions, 16 "Down" directions, 15 "left" directions and 19" right" directions are correctly detected from the original image. As shown in Table 2, the multi-patterns detection has a lower successful rate than those of single pattern detection.

Figure 10. Example images of our indoor signage dataset.

Table 1. Accuracy of single pattern detection

	Total number	Success number	Success rate(%)
Women	20	18	90
Men	20	16	80
Disable	10	9	90
Close	20	17	85
Open	20	17	85
Average	80	68	85

Table 2. Accuracy of multi patterns detection

	Total number	Success Number	Fail(to other types)	Success rate (%)
Women	20	14	2	70
Men	20	16	1	80
Up	20	18	0	90
Down	20	16	2	80
Right	22	19	1	86.3
Left	20	15	2	75
Average	122	98	8	80.3

The wrong signage detections fall into the following two categories. 1) When we build the saliency map, the query pattern will be ignored because of low resolution of the original images. 2) The bipartite graph matching method is sensitive to the resolution, perspective projection, and angle of the camera views. It is difficult to distinguish different types of signage and the success rate of the detection decreases if the image resolution is too low or captured with an extreme camera view. As shown in Figure 11, the "close" signage is not detected as attended area in the saliency map due to the low image resolution. Figure 12 demonstrates that "Men" signage is correctly detected by the saliency map, but is missed by the bipartite graph matching method.

(a) (b)

Figure 11. "Open" button of elevator: (a) Original image; (b) Saliency map

(a) (b) (c)

Figure 12. "Men" restroom: (a) Original image; (b) Saliency map; (c) Detection result

Table 3. Effectiveness of Saliency Map for Indoor Signage Detection

	Total	BGM only	SM and BGM
Men	20	11	14
Women	20	12	16
Total	40	23	30
Accuracy		57.5%	75%

2.4.4. Effectiveness of Saliency Map

To evaluate the effectiveness of a saliency map for the indoor signage detection, we compare the detection results with and without using saliency map detection by using the "Women's" and "Men's" restroom signage. As shown in Table 3, only 11"Men's" and 12 "Women's" restroom signs are correctly detected by applying bipartite graph matching on the image without performing saliency map based attended area detection. However, with saliency maps, we correctly detected 14 "Men's" and 16 "Women" restroom signs. The accuracy is increased from 57.5% (without using saliency maps) to 75% (with saliency maps). Saliency map can effectively eliminate disturbing information which would decrease the accuracy of the bipartite graph matching method.

2.4.5. Computation Cost Reduction

We further evaluate the computation time of the proposed method by comparing it to the traditional method of sliding window. The computation time for indoor signage detection and recognition is listed in Table 4. Our proposed method reduces computational cost by about 20% compared to the traditional method. This ensures real-time processing for developing navigation and wayfinding systems to help blind users.

Table 4. Computation Time Comparison (s)

Experiment	Traditional	Our's	Ratio
Button Detection (Open)	0.3247	0.2727	84.00%
Button Detection (Close)	0.3194	0.2760	86.40%
"Men" signage	0.4119	0.3472	78.46%
"Women" signage	0.2711	0.2319	85.55%

2.5. Summary

In this paper, we have proposed a novel object detection method for indoor signage recognition. The approach employs both saliency maps and bipartite graph matching. The system's ability to recognize elevator buttons and restroom signage demonstrates the effectiveness and efficiency of the proposed method.

Our future work will focus on extending our object detection method to handle larger changes of perspective projection, scale, point view, etc. We will also develop a prototype system of indoor signage detection and address the significant human interface issues associated with way finding for blind users.

Chapter 3: Camera-based Signage Detection and Recognition for Blind Persons

3.1. Introduction

In this paper, we propose a computer vision-based method for restroom signage detection and recognition. The proposed method contains both detection and recognition procedures. Detection procedure gets the location of a signage in the image. Recognition procedure is then performed to recognize the detected signage as 'Men", "Women", or "Disabled". The signage detection is based on effective shape segmentation, which is widely employed and achieved great success in traffic signage and traffic light detection [56]. The signage recognition employs SIFT feature-based matching, which is robust to variations of scale, translation and rotation, meanwhile partially invariant to illumination changes and 3D affine transformation.

Our proposed method in this paper is one component of a computer-vision based wayfinding and navigation aid for blind persons consists of a camera, a computer, and an auditory output device. For example, visual information would be captured via a mini-camera mounted on a cap or sunglasses, while image processing and speech output would be provided by a computer (with speech output via a BlueTooth earpiece). The recognition results can be presented to blind users by auditory signals (e.g., speech or sound).

This chapter is organized as following: Section 3.2 describes the method of this paper, which includes 3 steps: first, we do the image preprocessing, then we detect the head part and body part respectively, finally we employ the SIFT features to do the matching process to get the final result which contains the position information and what gender the signage indicates. Section 3.3 is the experiment part, it approves that our method is efficiency and effective. Section 3.4 is our conclusion and our future work.

3.2. Methodology for Restroom Signage Detection and Recognition

3.2.1. Method Overview

The proposed restroom signage recognition algorithm includes three main steps: image preprocessing, signage detection, and signage recognition as shown in Figure 13. Image preprocessing involves scale normalization, monochrome, binarization, and connected component labeling. Signage detection includes rule-based shape detection by detecting head and body parts of the signage respectively. Finally, the characteristic of restroom signage (e.g., for "Men", "Women", or "Disabled") is recognized by SIFT feature based matching distance between the detected signage region and restroom signage templates.

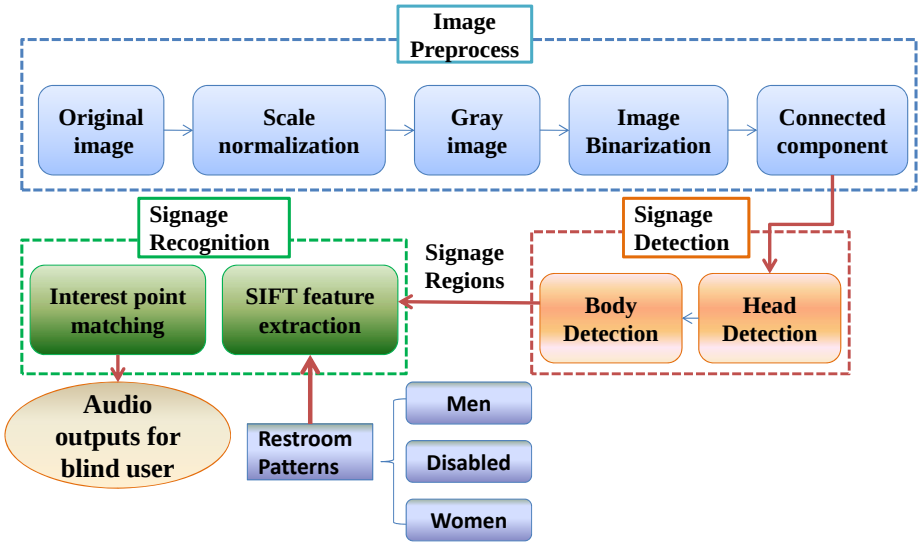


Figure 13. Flowchart of the proposed method

3.2.2. Image Preprocessing for Signage Detection

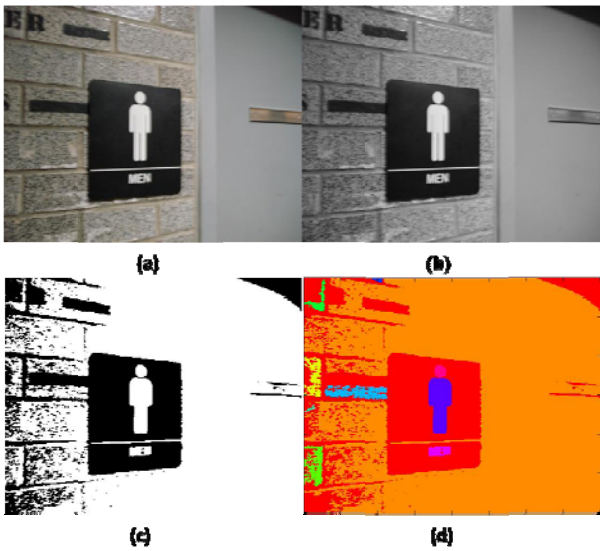


Figure 14. Image Preprocessing for restroom signage detection. (a) Original image, (b) Gray image; (c) Binary image; (d) Labeled connected components.

To effectively detect signage from an image, an image preprocessing is first conducted which includes three main steps: 1) convert input image to gray image; 2) binarize gray image to a binary image; and 3) perform connected component processing on binary images to find the connected pixels and eliminate small noises.

3.2.3. Signage Detection based on Shape and Compactness

We observe that most of images are upright and with relative stable illumination. Most important, the shape of the restroom signage in USA does not change much, which involves a circle-shaped "head" part and a more complicated "body" part as shown in Figure 14. In this section, we describe an effective rule-based method to locate restroom signage in images using shape information.

3.2.3.1. Detecting Head Part of a Restroom Signage

As shown in Figure 1, the restroom signage of all "Men", "Women", and "Disabled" has a circle-shaped head part. The most popular circle detection method is Hough transform. However, Hough transform detects circles by voting procedures based on *a-b-R* space [52]. Suppose processing a 200-by-200 image, the size *a-b-R* space is $200*200*100=4*10^6$, which is a large burden for the computers [56]. Meanwhile, Hough transform accept open circles, which do not represent the head part (closed circles), causing unpredicted recognition results. Thus we detect circles via the properties of connected components.

For each connected component which has a circle shape, the ratio of its perimeter and area is expected to be approximate to 4π. We set the rule as:

$$\text{If} \quad \alpha_2 \le \frac{\text{CC.peri\^{}2}}{\text{CC.Area}} \le \alpha_1 \text{, then the CC is } \textbf{Head}.$$

where CC.Area is the area of the connected component and CC.Peri is the perimeter of the connected component.

3.2.3.2. Detecting Body Part of a Restroom Signage

The body part of a restroom signage has more complicated shape which cannot be directly detected by simple shape detection method. Therefore, we detect a connected component if a body part based on the positions of the body part and head part of a restroom signage. A connected component is a body part if:

$$\beta_2 \le \frac{\text{CC.Area}}{\text{Head.Area}} \le \beta_1 \quad \& \quad \delta_2 \le \frac{\text{CC.Peri}}{\text{Head.Peri}} \le \delta_1 \quad \& \text{ CC is the nearest to the Head.}$$

where CC.Area and CC.Peri are the area and the perimeter of the connected component which is nearest the detected head part. All the parameters in the above equations are set by training of good quality sampled images from a restroom image database.

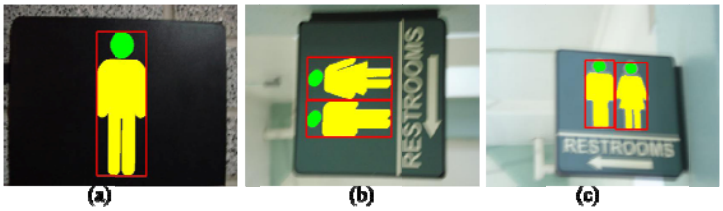


Figure 15. Example results of signage detection.

Figure 15 shows some example results of restroom signage detection. The green components indicate the detected "head" part, and yellow components indicate the "body" part, and red boxes show the signage locations in images which will be used for recognition.

3.2.4. Signage Recognition Based on SIFT Matching

3.2.4.1. SIFT Feature Extraction and Representation

SIFT features have been widely employed for object detection and recognition due to the robustness to variations of scale, translation, rotation, illumination, and 3D affine transformation. In order to perform signage recognition, we employ SIFT features and descriptors. SIFT feature extraction and representation contains two phases: (1) detect interest feature points and (2) feature point descriptor.

First, potential feature points are detected by searching overall scales and image locations through a difference-of-Gaussian (DoG) function pyramid. The DoG is a close approximation to the scale-normalized Laplacian-of-Gaussian to find the most stable image features [53] [54]. Hence, the locations of the points correspond to these most stable features are identified as interest feature points.

Second, the feature descriptor is created for each interest point by sampling the magnitudes and orientations of image gradients in a 16x16 neighbor region. The region is centered at the location of the interest point, rotated on the basis of its dominant gradient orientation and scaled to an appropriate size, and evenly partitioned into 16 sub-regions of 4x4 pixels. For each sub-region, SIFT accumulates the gradients of all pixels to orientation histograms with eight bins [55]. A 4x4 array of histograms, each with eight orientation bins, captures the rough spatial structure of the neighboring region. This 128-element vector, i.e. the feature descriptor for each interest point, is then normalized to unit length. Figure 4 shows the SIFT features extracted from "Women", "Men", and "Disabled" patterns. The detected interest points are the centers of green circles, and the radius of the green circle indicates the scales.

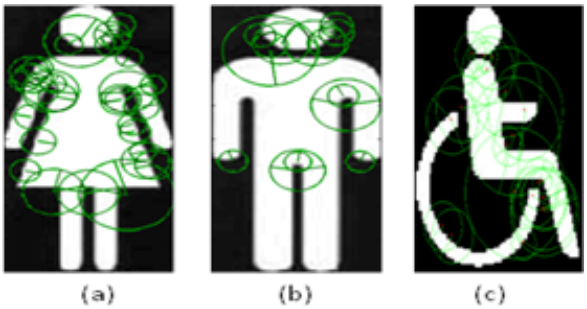


Figure 16. Examples of SIFT feature extraction for different restroom signage (a) “Women”; (b) “Men”; and (c) “Disabled”. The center of each green circle indicates one detected interest point, and the radius of the green circle indicates the scales.

3.2.4.2. Signage Recognition by SIFT Matching

In order to recognize the detected signage, SIFT-based interest points are first extracted from the template images of restroom signage patterns which are stored in a database (see Figure 16). Then, the features of the image region of the detected signage will be matched with those from the template signage patterns based on nearest Euclidean distance of their feature vectors. From the full set of matches, subsets of key points that agree on the object and its location, scale, and orientation in the new image are identified to filter out good matches. If two or more feature points in another image match a single point in the image, we assign the pair as the best match. In our method, two criteria are required for matching points (1) similar descriptors for corresponding feature; and (2) uniqueness for the correspondence.

Provided the number of matches between the signage template images and detected signage, the signage gets the maxima feature matches are selected as the most possible pattern. Figure 17 shows the matched features between the template signage patterns (the first row) and the detected signage regions (the second row).

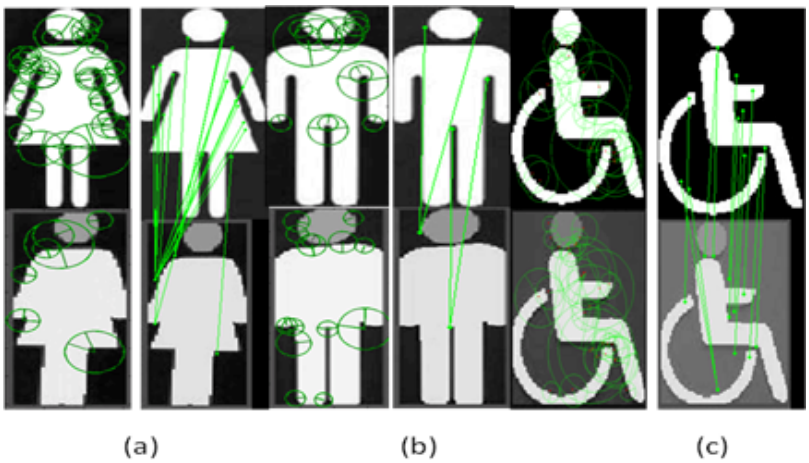


Figure 17. Examples of matched features between templates of restroom signage patterns (upper row) and the detected signage image regions (lower row).

3.3. Experimental Results

To validate the effectiveness and efficiency of our method, we have collected a database which contains 96 images of restroom signage including patterns of “Women”, “Men”, and “Disabled”. There are total 50 “Men” signage, 42 “Women” signage, and 10 images of “Disabled” signage. As shown in Figure 6, the database includes the changes of illuminations, scale, rotation, camera view, perspective projection, etc. Some of the images contain both signage of “Men” and “Women”, or both signage of “Men” and “Disabled”, or “Women” and “Disabled”.

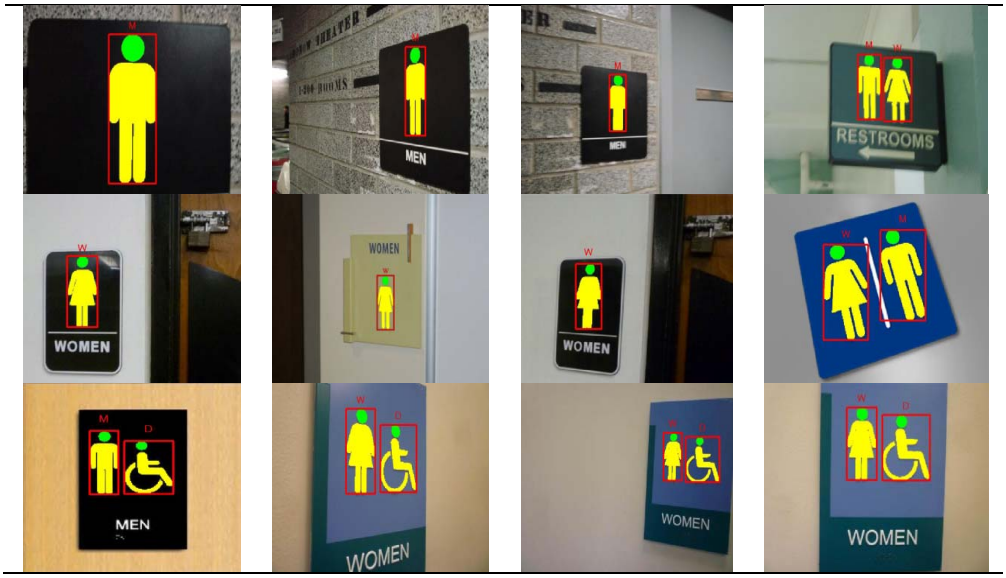


Figure 18. Sample images with signage detection and recognition in our database include changes of illuminations, scale, rotation, camera view, and perspective projection, etc. The red boxes show the detected signage region, while the letter above each red box indicates the recognition of the signage: "W" for "Women", "M" for "Men", and "D" for "Disabled".

Figure 19. Sample signage images of each step of the proposed method. Columns from left to right: original images, binarized images, images with connected components, and detected and recognized signage.

Our method can handle signage with variations of illuminations, scales, rotations, camera views, perspective projections. We evaluate the recognition accuracy of the proposed method.

As shown in Table 5, the proposed algorithm achieves accuracy of detection rate 89.2% and of recognition rate 84.3% which correctly detected 91 and recognized 86 signage of total`102 signage in our dataset. Some examples of the detected restroom signage from different environments are shown in Figure 18 and Figure 19. The red boxes show the detected signage region, while the letter above each red box indicates the recognition of the signage: "W" for "Women", "M" for "Men", and "D" for "Disabled".

Table 5. Restroom signage recognition accuracy

	Men(50)	Women(42)	Disabled(10)
Men	41	3	0
Women	2	36	0
Disabled	0	0	9

Figure 20 demonstrates several signage examples which our method fails to detect and recognize. We observe that the failures are caused by the following three reasons: 1) large camera view changes which can cause large shape distortion; 2) image blurry due to camera motion; and 3) low image resolution when the user is far from the signage

Figure 20. Examples of Failures

We further verify the computation time of the proposed method. The experiments are carried on a computer with a 2GBHz processor and 1GB memory. The proposed algorithm is implemented in Matlab code. The average time for detecting and recognizing signage from 30 testing images is 0.192s. This ensures real-time processing for developing navigation and wayfinding systems to help blind and vision impaired users.

3.4. Summary

To assist blind persons independently accessing unfamiliar environments, we have proposed a novel method to detect and recognize restroom signage based on both shape and appearance features. The proposed method can handle restroom signage with variations of scales, camera views, perspective projections, and rotations. The experiment results demonstrate the effectiveness and efficiency of our method. Our future work will focus on detecting and recognizing more types of signage and incorporating context information to improve indoor navigation and wayfinding for blind people. We will also address the significant human interface issues including auditory displays and spatial updating of object location, orientation, and distance.

Chapter 4: Detecting Stairs and Pedestrian Crosswalks for the Blind by RGBD Camera

4.1. Introduction

In this paper, we propose a computer vision-based method to detect stair-cases and pedestrian crosswalks by using a commodity RGBD camera. The recent introduction of the cost-effective RGBD cameras eases the task by providing both RGB information and depth information of the scene. As shown in Figure 1, our method consists of three main steps. First, a group of parallel lines are detected via Hough transform and line fitting with geometric constraints from RGB information (see details in Section 2.1). In order to distinguish stairs and pedestrian crosswalks, we extract the feature of one dimension depth information according to the direction of the detected longest line from the depth image. Then the feature of one dimension depth information is employed as the input of a SVM-based classifier to recognize stairs and pedestrian crosswalks. For stairs, a further detection of upstairs and downstairs is conducted. Furthermore, we estimate the distance between the camera and stairs for the blind users.

The paper is organized as following: Section 4.2 describes the methodology of our proposed algorithm including 1) detection whether the scene image contains stair-cases or pedestrian crosswalks based on RGB image analysis; 2) since both stairs and pedestrian crosswalks are featured by a group of parallel lines in RGB images, we further employ depth information to distinguish stairs from pedestrian crosswalks, then stairs will be further recognized as upstairs and downstairs. Section 4.3 displays the evaluation effectiveness and efficiency of proposed method and summarizes the experiment results. Section 4.4 concludes the paper and our future work.

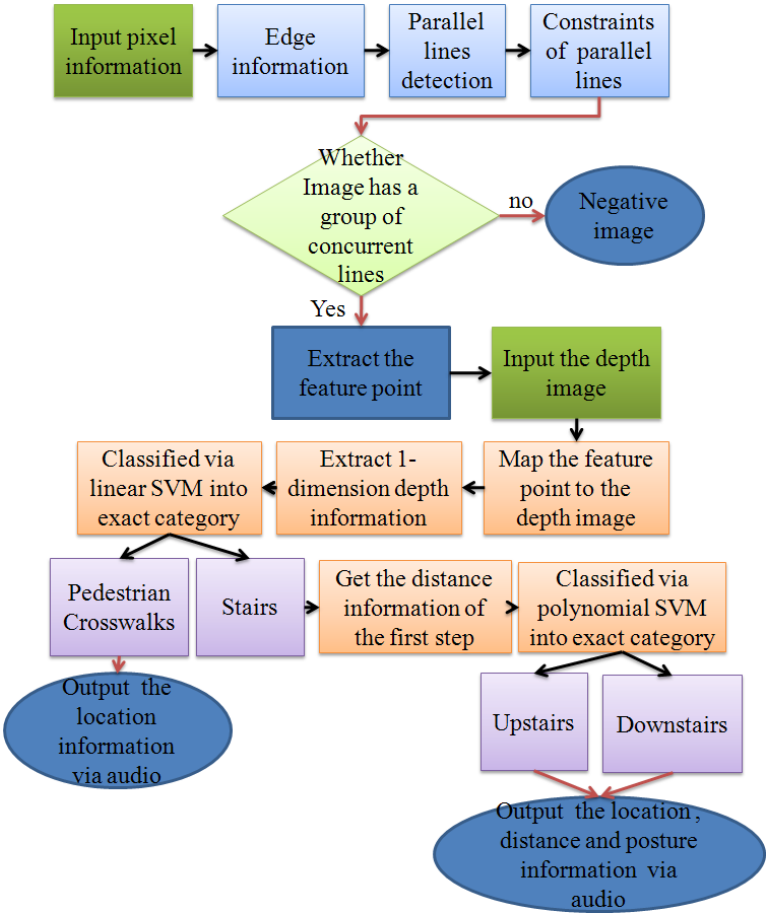


Figure 21. Flow chart of the proposed algorithm for stair and pedestrian crosswalk detection and recognition.

4.2. Methodology of RGBD Camera based Stair and Pedestrian Crosswalk Detection

4.2.1. Detecting Candidates of Pedestrian Crosswalks and Stairs from RGB images

There are various kinds of stair-cases and pedestrian crosswalks. In this paper, we focus on stair cases with uniform trend and steps, and pedestrian crosswalks of the most regular zebra crosswalks with alternating white bands. In our application of blind navigation and wayfinding, we focus on detecting stairs or pedestrian crosswalks in a close distance.

Stairs consists of a sequence of steps which can be regarded as a group of consecutive curb edges, and pedestrian crosswalks can be characterized as an alternating pattern of black and white stripes. To extract these features, we start with an edge detection to obtain the edge map from RGB image of the scene and then perform a Hough transform to extract the lines in the extracted edge map image. These lines for are parallel for both stairs and pedestrian crosswalks. Therefore, a group of concurrent parallel lines will most likely represent the structure of stairs and pedestrian crosswalks. In order to eliminate the noise from unrelated lines, we add constraints including the number of concurrent lines, line length, etc.

Extracting Parallel Lines based on Hough Transform: We apply Hough transform to detect straight lines based on the edge points. A number of edge points (x_i, y_i) in an image that form a line can be expressed in the slope-intercept form: $y=ax+b$, where a is the slope of the line and b is the y-intercept. The main idea here is to consider the characteristics of a straight line not as image points (x_1, y_1), (x_2, y_2), etc., but instead, in terms of its parameters. Based on that fact, the straight line $y=ax+b$ can be represented as a point (a, b) in the parameter space. However, we face the problem that vertical lines give rise to unbounded values of the parameters a and b. Considering the unbounded values of the parameters a and b, it is better to use the Polar Coordinates, denoted r and θ, for the lines in the Hough transform (as shown in Figure 2).

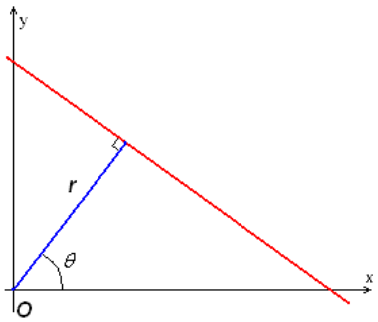


Figure 22. Illustration of polar coordinates of a line.

The parameter r represents the distance between the line and the origin, while θ is the angle of the vector from the origin to the closest point, then the equation of a line can be represented as:

$$r = y\sin\theta + x\cos\theta \qquad (11)$$

The algorithm of Hough transform line fitting is summarized as following:

Step1: Detect edge maps from the RGB image by edge detection.

Step2: Compute the Hough transform of the RGB image to obtain r and θ.

Step3: Calculate the peaks in the Hough transform matrix.

Step4: Extract lines in the RGB image.

Step5: Detect a group of parallel lines based on constraints such as the length and total number of detected lines of stairs and pedestrian crosswalks.

As shown in Figure 23(c), Figure 24(c), and Figure 25(c), the detected parallel lines of stairs and pedestrian crosswalks are marked as green, while yellow dots and red dots represent the beginning and the end of the lines respectively. However, these lines are often separated with small gaps caused by noises, so we group the line fragments as the same line if the gap less than a threshold. In general, stairs and pedestrian crosswalks contain multiple parallel lines with a reasonable length. If the length of a line $\leq \phi$, then the line is not belong to the line group. And if the number of parallel lines less than β, the scene image is a negative image which does not contain stairs and pedestrian crosswalks. In our experiment, we set the line length ϕ as 60 pixels in the acquired images and the number parallel lines β as 5.

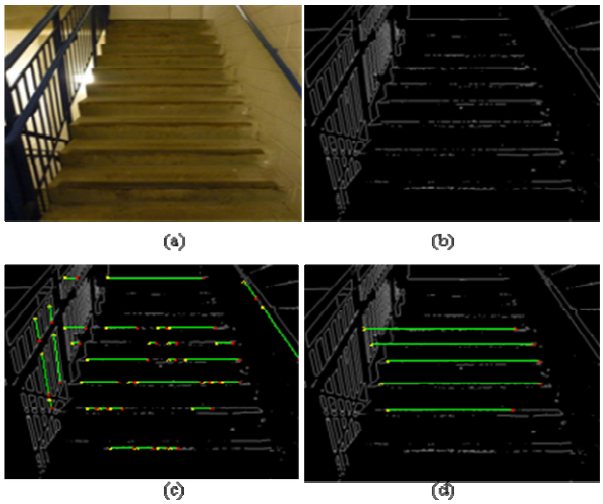


Figure 23. An example of upstairs. (a) Original image; (b) edge detection; (c) line detection; (d) concurrent parallel lines detection (yellow dots represent the beginnings, red dots represent the ends of the lines, and green lines represent the detected lines.)

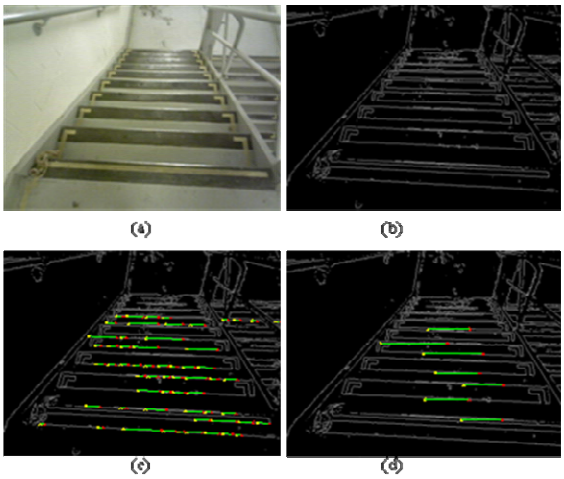


Figure 24. An example of downstairs. (a) Original image; (b) edge detection; (c) line detection; (d) concurrent parallel lines detection (yellow dots represent the beginnings, red dots represent the ends of the lines, and green lines represent the detected lines.)

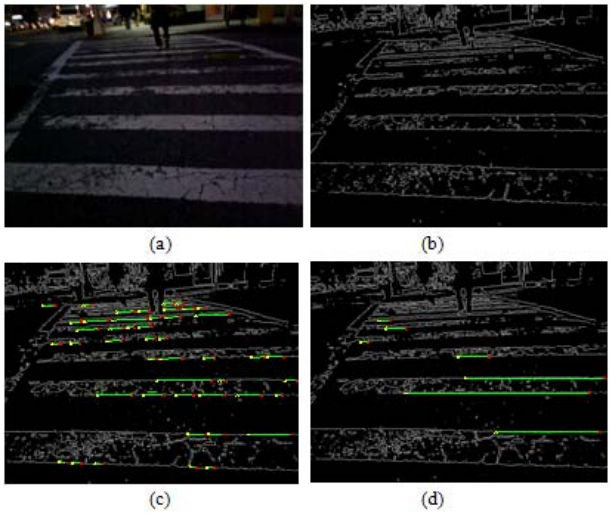


Figure 25. An example of Pedestrian crosswalks. (a) Original image; (b) edge detection; (c) line detection; (d) concurrent parallel lines detection (yellow dots represent the beginnings, red dots represent the ends of the lines, and green lines represent the detected lines.)

4.2.2. Recognizing Pedestrian Crosswalks and Stairs from Depth Images

Based on the above algorithm, we can detect the candidates of stairs or pedestrian crosswalks by detecting parallel lines with constraint condition in a scene image captured by a RGBD camera. From the depth images, we observe that upstairs have rising steps and downstairs have decreasing step, and pedestrian crosswalks are flat with smooth depth change as shown in Figure 26. Considering the safeness for the visually impaired people, and the further application for the robotic, it is necessary to classify the different stairs and pedestrian crosswalks into the correct categories.

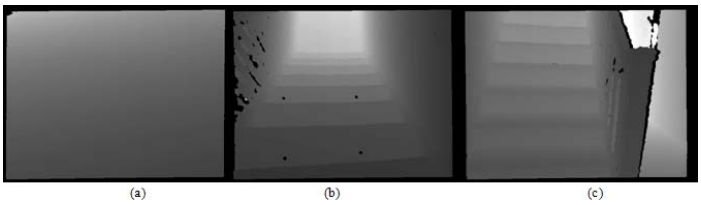


Figure 26. Depth images of (a) pedestrian crosswalks, (b) downstairs, and (c) upstairs.

In order to distinguish stairs and pedestrian crosswalks, we first calculate the orientation and position for extract the one-dimensional based feature from depth information. As shown in Figure 27, the orientation is perpendicular to the parallel lines detected from RGB images. The position will be determined by the middle point of the longest line of the parallel lines. In Figure 27, the blue square indicates the middle point of the longest line and the red line shows the orientation to calculate the one-dimensional depth features. The typical one-dimensional depth feature for upstairs, downstairs, and pedestrian crosswalks are demonstrated in Figure 28.

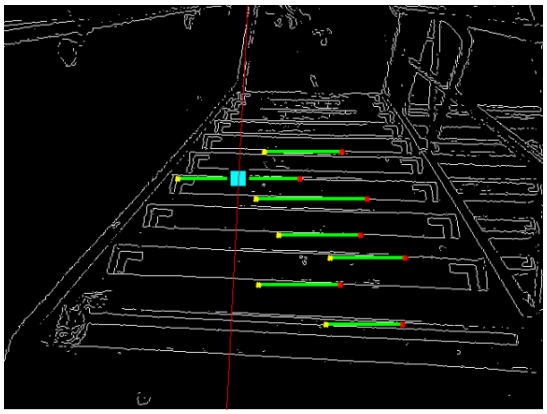


Figure 27. Orientation and position to calculate one-dimensional depth features from edge image. The blue square indicates the middle point of the longest line and the red line shows the orientation which is perpendicular to the detected parallel lines.

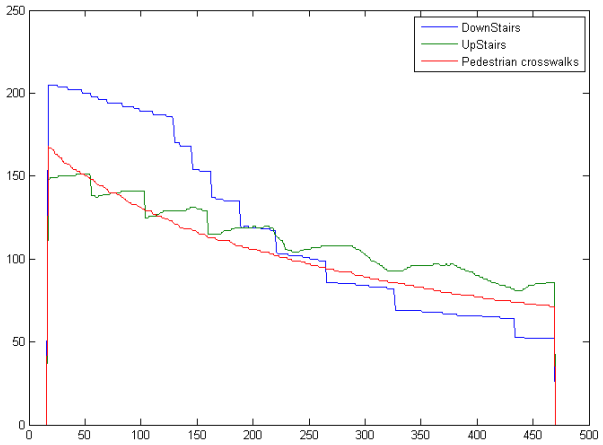


Figure 28. One-dimensional depth feature for upstairs (green), downstairs (blue), and pedestrian crosswalks (red). The horizontal axis indicates the distance from the camera in centimeters. The vertical axis represents the intensity of the depth image.

As shown in Figure 26, the resolution of depth images is 480*640 pixels. The effective depth range of the RGBD camera is about 0.15 to 4.0 meters. The intensity value range of the depth images is [0, 255]. Therefore, as shown in Figure 8, the intensities of all the curves of the one dimension depth features are between 50 and 220 but are 0 if the distance is out of the depth range of a RGBD camera.

In order to classify upstairs, downstairs, and pedestrian crosswalks, we propose a hierarchical SVM structure by using the extracted one-dimensional depth features. The SVM builds a set of hyper-planes in an infinite-dimensional space, which can be used for classification, regression, or other tasks. The high classification accuracy can be achieved by the hyper-plane that has the largest distance to the nearest training data point of any class. In the classification section, we will have two steps, first classify pedestrian crosswalks from stairs, and then we further classify upstairs and downstairs.

4.2.3. Estimating Distance between Stairs and the Camera

When walking on stairs, we should adjust our foot height as the stairs has a steep rising or decreasing. For blind users, stairs, in particular downstairs, may cause injury if they fall. Therefore, it is essential to provide the distance information

to the blind or visually impaired individuals how far is the first step of the stairs away from the camera position to remind them when they should adjust their foot height. In our method, the distance information between the first step of the stairs and the camera position will be calculated by detecting the first turning point from the one-dimension depth information as shown in Figure 29 marked as the red dots.

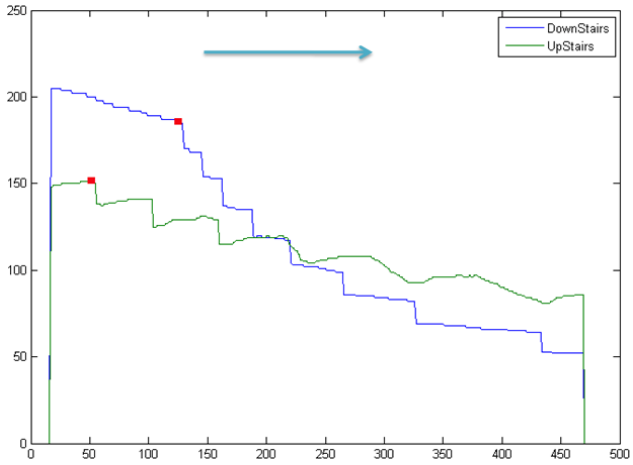


Figure 29. Detecting the first turning points (red points) of the one-dimensional depth features of upstairs and downstairs.

From the near distance to far distance (e.g., from left side to the right side as the blue line with arrow shown in Figure 29) along the one-dimensional depth features, a point x satisfies the following two conditions is considered as a turning point:

$$f(x) - f(x-1) > \lambda \qquad (12)$$

or

$$f'(x) - f'(x-1) > \varepsilon \qquad (13)$$

where f(x) is the intensity value of the depth information, λ and ε are the thresholds. In our experiment, we set λ =8 and ε =50.

After we obtain the position of the turning point which indicates the first step of the stairs, the distance information from the camera and the first step of the stairs can be read from the original RGBD depth data. This distance will be provided to the blind traveler by speech.

4.3. Experiments and Discussion

4.3.1. Database

To evaluate the effectiveness and efficiency of the proposed method, we collect two databases: a testing database and a training database. The testing database contains 106 stairs including 56 upstairs and 50 downstairs, 52 pedestrian crosswalks, and 70 negative images which contain neither stairs nor pedestrian crosswalks. Some of negative images contain objects structured with a group of parallel lines such as bookshelves. The training database contains 30 images for each category to train the SVM classifiers. The images in the databases include small changes of camera view angels

$[-30^o, 30^o]$ because the visually impaired people pay more attention to the area in front of them. The experiment example used in our algorithm is shown in Figure 10. The first row displays examples of upstairs with different camera angles and the second row shows the corresponding depth images. Similarly, the third and fourth rows are the RGB depth images for examples of downstairs, and the fifth and sixth rows are the examples of pedestrian crosswalks.

4.3.2. Experimental Results

We evaluate the accuracy of the detection and the classification of our proposed method. The proposed algorithm achieves an accuracy of detection rate at 91.14% among the positive image samples and 0% false positive rate as shown in Table 6. For the detection step, we correctly detect 103 stairs from 106 images, and 41 pedestrian crosswalks from 52 images of pedestrian crosswalks. Here, positive image samples indicate images containing either stairs or pedestrian crosswalks, and negative image samples indicates images containing neither stairs nor pedestrian crosswalks.

The negative samples include some objects such as bookshelves, which are constructed similar edges as stairs and pedestrian crosswalks as shown in Figure 31. With the current camera configuration, in general, only one to two shelves can be captured. The detected parallel lines will not meet the constraint conditions as described in Section 4.2.1. Therefore, the bookshelves will not be detected as candidates of stairs and pedestrian crosswalks.

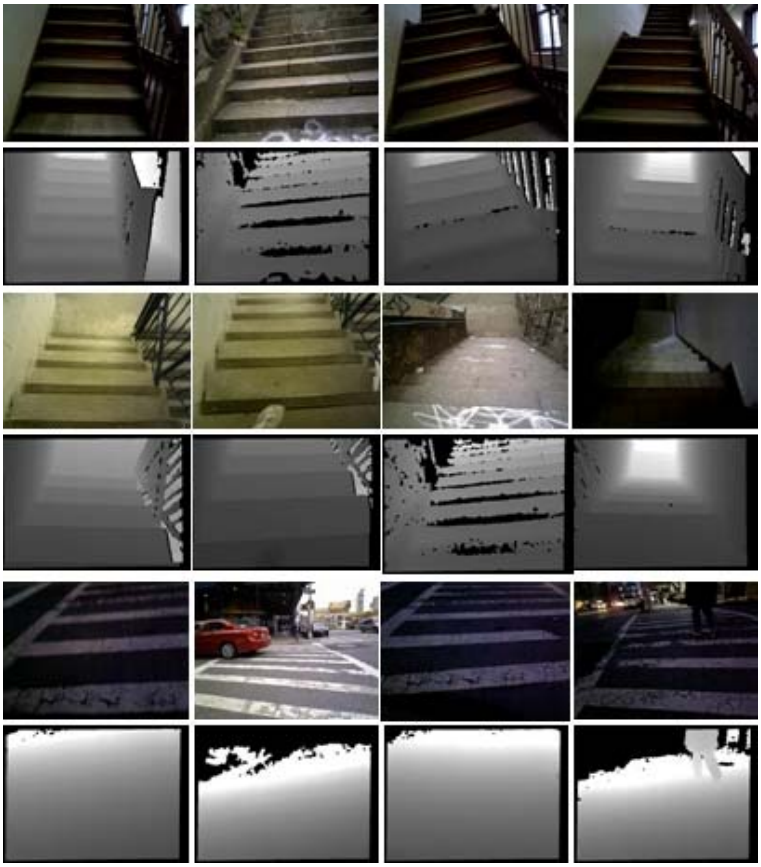


Figure 30. Examples of RGB and depth images for upstairs (1st and 2nd rows), downstairs (3rd and 4th rows), and pedestrian crosswalks (5th and 6th rows) in our database.

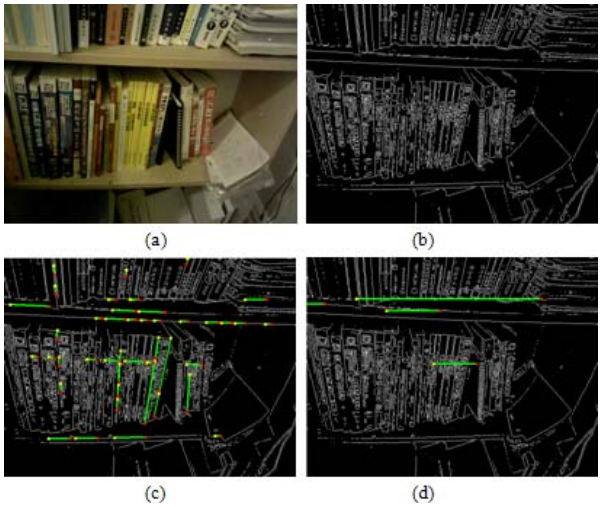


Figure 31. Negative examples of a bookshelf which has similar edge lines to stairs and pedestrian crosswalks.

In order to classify stairs and pedestrian crosswalks, the detected positive images are input into a SVM-based classifier. As shown in Table 7, our method achieves a classification rate for the stairs and pedestrian crosswalks at 95.8% which correctly classified 138 images from 144 detected candidates. Total of 6 images of stairs are wrongly classified as pedestrian crosswalks.

For stairs, we further classify they are upstairs or downstairs by inputting the one-dimensional depth features into a different SVM classifier. We achieve an accuracy rate of 90.2%. More details of the classification of upstairs and downstairs are listed in Table 8

Table 6. Detection accuracy of stairs and pedestrian crosswalks

Classes	No. of Samples	Correctly Detected	Detection Accuracy
Stairs	106	103	97.2%
Crosswalks	52	41	78.9%
Negative samples	70	70	100%
Average	228	214	93.9%

Table 7. Accuracy of classification between stairs and pedestrian crosswalks

	Stairs	Crosswalks
Stairs	97	0
Crosswalks	6	41

Table 8. Accuracy of classification between upstairs and downstairs

	Upstairs	Downstairs
Upstairs	48	5
Downstairs	5	45

In database capture, we observe that it is hard to capture good quality depth images of pedestrian crosswalks compared to capture images of stairs. The main reason is the current RGBD cameras cannot obtain good depth information for outdoor scenes if the sunshine is too bright. Therefore, the field of view of the obtained depth maps is restricted compared to the RGB images. Some of the images our method cannot handle are shown in Figure 32. For example, the depth information of some parts of the images (see the 2nd and 4th columns of the 6th row of Figure 32) is missing. Furthermore, the white band patterns of pedestrian crosswalks are often disappeared caused by the long time exposure and no well maintained as shown in Figure 31(c). In this case, it will be hard to extract parallel lines to satisfy the candidate detection constraints we described in Section 4.2.1. In our method, stairs with less than 3 steps will not be able to detected, as shown in Figure 32(a) and (d).

Figure 32. Examples of our proposed method fails. (a) Downstairs with poor illumination; (b) Upstairs with less detected lines caused by noise; (c) Pedestrian crosswalks with missing white patterns; and (d) Stairs with less steps.

4.4. Summary

We have developed a novel method for automatic detection of pedestrian crosswalks, upstairs, and downstairs by using a RGBD camera to improve the travel safeness of the blind and visually impaired people. The proposed method can run in real time. Our method has been evaluated on the database of stairs and pedestrian crosswalks, and achieved accuracy rates of 91.1% for detection stairs and pedestrian crosswalks from scene images, 95.8% for classification of stairs and pedestrian crosswalks, and 90.3% for classification of upstairs and downstairs,

Our further research will focus on enhancing our algorithm to handle stairs and pedestrian crosswalks with large perspective projections, more types of objects, user interface study with evaluation by blind subjects.

Chapter 5: Conclusion and Future Work

In this thesis, we have proposed different methods of object detection and recognition to help blind people or visually impaired to access unfamiliar environment independently.

In the Chapter 2, we employed both saliency maps and bipartite matching to recognize elevator buttons, restroom signage and emergency exit directions.

In the Chapter 3, we have proposed a novel method to detect and recognize restroom signage based on both shape and appearance features. The proposed method can handle restroom signage with variations of scales, camera views, perspective projections, and rotations.

In the Chapter 4, we have developed a novel method for automatic detection of pedestrian crosswalks, upstairs, and downstairs by using a RGBD camera to improve the travel safeness of the blind and visually impaired people. The proposed method can run in real time. Our method has been evaluated on the database of stairs and pedestrian crosswalks, and achieved accuracy rates of 91.1% for detection stairs and pedestrian crosswalks from scene images, 95.8% for classification of stairs and pedestrian crosswalks, and 90.3% for classification of upstairs and downstairs,

Our future work will focus on optimize the algorithms to improve the computing efficiency, design user interface to study with blind and visually impaired subjects. Furthermore, we should test more types of objects to make our algorithm widely used. It is also essential to make more experiments to further approve our method.

References:

[1]. Kocur, R. Parajasegaram, and G. Pokharel, Global Data on Visual Impairment in the Year 2002. *Bulletin of the World Health Organization*, 82, pp.844-851, 2004.

[2]. A. Baker, Blind Man is Found Dead in Elevator Shaft, The New York Times, City Room, May 2010.

[3]. A. Neves, A. Pinho, D. Martins, and B. Cunha, "An efficient omnidirectional vision system for soccer robots: From calibration to object detection," Mechatronics, March 2011. Vol. 21, No. 2, pp. 399-410.

[4]. G. Kreiman, 2008, "Biological object recognition," Scholarpedia, 3(6):2667. http://www.scholarpedia.org /article /Biological_object_recognition.

[5]. Biederman, I. 1987, "Recognition-by-Components: A Theory of Human Image Understanding," Psychological Rev., vol. 94, pp. 115-147, 1987.

[6]. M. Potter and E. Levy, Recognition memory for a rapid sequence of pictures. Journal of Experimental Psychology, 1969. 81: 10-15.

[7]. S. Thorpe, D. Fize, and C. Marlot, Speed of processing in the human visual system. Nature, 1996. 381: 520-522.

[8]. C. Hung, G. Kreiman, T. Poggio, and J. DiCarlo, Fast Read-out of Object Identity from Macaque Inferior Temporal Cortex. Science, 2005. 310: 863-866.

[9]. H. Schneiderman and T. Kanade. A statistical approach to 3d object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00)*, 2000.

[10]. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'01)*, 2001.

[11]. M. Weber, M. Welling, and P. Perona, "Unsupervised Learning of Models for Recognition," Proc. European Conf. Computer Vision, vol. 2, pp. 101-108, 2000.

[12]. Seeing with Sound – The vOICe, http://www.seeingwithsound.com/

[13]. The Smith-Kettlewell Rehabilitation Engineering Research Center (RERC) develops new technology and methods for understanding, assessment and rehabilitation of blindness and visual impairment. http://www.ski.org/Rehab/

[14]. V. Ivanchenko, J. Coughlan and H. Shen. "Crosswatch: a Camera Phone System for Orienting Visually Impaired Pedestrians at Traffic Intersections." 11th International Conference on Computers Helping People with Special Needs (ICCHP '08). 2008.

[15]. R. Manduchi, J. Coughlan and V. Ivanchenko. "Search Strategies of Visually Impaired Persons using a Camera Phone Wayfinding System." 11th International Conference on Computers Helping People with Special Needs (ICCHP '08). 2008

[16]. V. Ivanchenko, J. Coughlan and H. Shen. "Detecting and Locating Crosswalks using a Camera Phone." Fourth IEEE Workshop on Embedded Computer Vision, in conjunction with Computer Vision and Pattern Recognition (CVPR '08). 2008

[17]. H. Shen and J. Coughlan. "Grouping Using Factor Graphs: an Approach for Finding Text with a Camera Phone." Workshop on Graph-based Representations in Pattern Recognition, 2007.

[18]. A. Zandifar, R. Duraiswami, A. Chahine, L. Davis, "A video based interface to textual information for the visually impaired". In: Proc. IEEE 4th international conference on multimodal interfaces, 2002.

[19]. M. Everingham, B. Thomas, and T. Troscianko, "Wearable Mobility Aid for Low Vision Using Scene Classification in a Markov Random Field Model Framework," International Journal of Human Computer Interaction, Volume 15, Issue 2, 2003.

[20]. S. Shoval, I. Ulrich, and J. Borenstein, "Computerized Obstacle Avoidance Systems for the Blind and Visually Impaired." Invited chapter in "Intelligent Systems and Technologies in Rehabilitation Engineering." Editors: Teodorescu, H.N.L. and Jain, L.C., CRC Press, ISBN/ISSN: 0849301408, Publication Date: 12/26/00, pp. 414-448.

[21]. V. Pradeep, G. Medioni, and J. Weiland, Piecewise Planar Modeling for Step Detection using Stereo Vision, Workshop on Computer Vision Applications for the Visually Impaired, 2008.

[22]. Hasanuzzaman, F., Yang, X., and Tian, T., Robust and Effective Component-based Banknote Recognition for the Blind, IEEE Transactions on Systems, Man, and Cybernetics--Part C: Applications and Reviews, Vol. 41, Issue. 5, 2011. 10.1109/TSMCC.2011.2178120

[23]. X. Yang, S. Yuan, and Y. Tian, Recognizing Clothes Patterns for Blind People by Confidence Margin based Feature Combination, International Conference on ACM Multimedia, 2011.

[24]. S. Yuan, Y. Tian, and A. Arditi, "Clothing Matching for Visually Impaired Persons", Technology and Disability 23, Page1-11, 2011. The online version if the article is available at: http://dx.doi.org/10.3233/TAD-2011-0313.

[25]. C. Yi and Y. Tian. Assistive Text Reading from Complex Background for Blind Persons, The 4th International Workshop on Camera-Based Document Analysis and Recognition (CBDAR), 2011.

[26]. C. Yi and Y. Tian. Text Detection in Natural Scene Images by Stroke Gabor Words, The 11th International Conference on Document Analysis and Recognition (ICDAR), 2011.

[27]. C. Yi and Y. Tian, "Text String Detection from Natural Scenes by Structure-based Partition and Grouping", IEEE Transactions on Image Processing, Vol. 20, Issue 9, 2011. PMID: 21411405.

[28]. Wang, S. H., Tian, Y. L. Indoor signage detection based on saliency map and Bipartite Graph matching, IEEE International Conference on Bioinformatics and Biomedicine Workshops (2011)

[29]. X. Yang, Y. Tian, C. Yi, and A. Arditi, Context-based Indoor Object Detection as an Aid to Blind Persons Accessing Unfamiliar Environment", International Conference on ACM Multimedia, 2010.

[30]. X. Yang and Y. Tian, "Robust Door Detection in Unfamiliar Environments by Combining Edge and Corner Features", 3rd Workshop on Computer Vision Applications for the Visually Impaired (CVAVI), 2010.

[31]. J. Coughlan and H.Y. Shen, A fast algorithm for finding crosswalks using figure-ground segmentation. In Proc. 2nd Workshop on Applications of Computer Vision, in conjunction with ECCV, 2006.

[32]. R. Advanyi, B. Varga, and K. Karacs, "Advanced crosswalk detection for the Bionic Eyeglass," 12th International Workshop on Cellular Nanoscale Networks and Their Applications (CNNA), pp.1-5, 2010

[33]. S. Se, "Zebra-crossing Detection for the Partially Sighted", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, pp. 211-217, 2000.]

[34]. S. Se and M. Brady, "Vision-based Detection of Stair-cases", Proceedings of Fourth Asian Conference on Computer Vision (ACCV), pp. 535-540, 2000.

[35]. Uddin and T. Shioyama, "Bipolarity and Projective Invariant-Based Zebra-Crossing Detection for the Visually Impaired," 1st IEEE Workshop on Computer Vision Applications for the Visually Impaired, 2005

[36]. L Lausser, F. Schwenker, and G Palm，"Detecting zebra crossings utilizing AdaBoost, European Symposium on Artificial Neural networks", Advances in computational intelligence and learning, 2008.

[37]. Y. Tian, X. Yang, C. Yi, and A. Arditi, "Toward a Computer Vision-based Wayfinding Aid for Blind Persons to Access Unfamiliar Indoor Environments," Machine Vision and Applications, 2012, DOI: 10.1007/s00138-012-0431-7.

[38]. W. Tsao, A. Lee, Y. Liu, T.-W. Chang, and H. -H. Lin, "A data mining approach to face detection," Pattern Recognition, March 2010, Vol. 43, No. 3, pp. 1039-1049.

[39]. S. Leyk, R. Boesch, and R. Weibel, "Saliency and semantic processing: Extracting forest cover from historical topographic maps," Pattern Recognition, May 2006, Vol. 39, No. 5, pp. 953-968.

[40]. H. Shi, and Y. Yang, "A computational model of visual attention based on saliency maps," Applied Mathematics and Computation, May 2007, Vol. 188, No. 2, pp. 1671-1677.

[41]. K. Makino, T. Takabatake, and S. Fujishige, "A simple matching algorithm for regular bipartite graphs," Information Processing Letters, Novermber 2002, Vol. 84, No. 4, pp. 189-193.

[42]. K. Riesen, and H. Bunke, "Approximate graph edit distance computation by means of bipartite graph matching," Image and Vision Computing, June 2009. Vol. 27, No. 7, pp. 950-959.

[43]. Y. D. Zhang, and L. N. Wu, "Pattern recognition via PCNN and Tsallis entropy," Sensors, 2008. Vol. 8, No. 11, pp. 7518-7529.

[44]. C. Davies, W. Tompkinson, N. Donnelly, L. Gordon, and K. Cave, "Visual saliency as an aid to updating digital maps," Computers in Human Behavior, July 2006. Vol. 22, No. 4, pp. 672-684.

[45]. S. J. Park, K. H. An, and M. Lee, "Saliency map model with adaptive masking based on independent component analysis," Neurocomputing, December2002, Vol. 49, No. 1-4, pp. 417-422.

[46]. A, Shokoufandeh, I. Marsic, and S. J. Dickinson, "View-based object recognition using saliency maps," Image and Vision Computing, April 1999. Vol. 17, No. 5-6, pp. 445-460.

[47]. C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map," Current Biology, 8. Novemember 2005, Vol. 15, No. 21, pp. 1943-1947.

[48]. F. Davignon, J. F. Deprez, and O. Basset, "A parametric imaging approach for the segmentation of ultrasound data," Ultrasonics, December 2005, Vol. 43, No. 10, pp. 789-801.

[49]. A. Noureldin, A. El-Shafie, and M. R Taha, "Optimizing neuro-fuzzy modules for data fusion of vehicular navigation systems using temporal cross-validation," Engineering Applications of Artificial Intelligence, February 2007. Vol. 20, No. 1, pp. 49-61

[50]. C. R. Pranesachar, "A class of matching-equivalent' bipartite graphs," Discrete Mathematics, May 1999. Vol. 203, No. 1-3, pp. 207-213.

[51]. N. Kakimura, "Matching structure of symmetric bipartite graphs and a generalization of Pólya's problem," Journal of Combinatorial Theory, Series B, Novemember 2010, Vol. 100, No. 6, pp. 650-670.

[52]. Cao, M.Y., Ye, C. H., Doessel, O., Liu. C., Spherical parameter detection based on hierarchical Hough transform, Pattern Recognition Letters, vol. 27, pp. 980-986 (2006)

[53]. Lindeberg, T., Scale-space theory: “A basic tool for analyzing structures at different scales”. J. Appl. Statist, vol. 21, 224–270 (2004)

[54]. Mikolajczyk, K., Schmid, C., An affine invariant interest point detector. In: Proc. European Conf. of Computer Vision (ECCV), pp. 128–142 (2002)

[55]. Matsui, Y., Miyoshi, Y., Difference-of-Gaussian-Like Characteristics for Optoelectronic Visual Sensor, Signal Processing & Analysis, vol. 7, pp. 1447 – 1452 (2007)

[56]. Omachi M, Omachi, S. Traffic light detection with color and edge information. Computer Science and Information Technology, 2nd IEEE International Conference in Beijing, pp.284-287 (2009)

Publications:

1. S. Wang and Y. Tian, Indoor signage detection based on saliency map and Bipartite Graph matching, International Workshop on Biomedical and Health Informatics, 2011.
2. S. Wang and Y. Tian, Camera-based Signage Detection and Recognition for Blind Persons, 13th International Conference on Computers Helping People with Special Needs (ICCHP), 2012.
3. S. Wang, C. Yi, and Y. Tian, "Signage Detection and Recognition for Blind Persons to Access Unfamiliar Environments," Accepted, Journal of Computer Vision and Image Processing, 2012.
4. S. Wang and Y. Tian, Detecting Stairs and Pedestrian Crosswalks for the Blind by RGBD Camera, Submitted, IEEE International Conference on Bioinformatics and Biomedicine (BIBM12), 2012.
5. S. Wang, X. Yang, and Y. Tian, Detecting Signage and Doors for Blind Navigation and Wayfinding, Submitted, Network Modeling Analysis in Health Informatics and Bioinformatics.